What traits do academics value in student writing? Insights from a psychometric approach

Neil Murray* and Gerard Sharpling

Centre for Applied Linguistics, University of Warwick, Coventry, UK

Neil Murray (Corresponding author)
Centre for Applied Linguistics, University of Warwick
Social Sciences Building
Warwick University
Coventry CV4 7AL
United Kingdom
Tel: 02476-524179
Email: N.L.Murray@warwick.ac.uk
ORCiD Identifier: orcid.org/0000-0002-5372-5529

Neil Murray is Associate Professor of Applied Linguistics at the Centre for Applied Linguistics, University of Warwick, UK, and adjunct member of the Research Centre for Languages & Cultures at the University of South Australia, where he was previously Head of Language and Literacy. He has published widely on language assessment and academic literacy and his research interests include English language policy and regulation in higher education, English as a medium of instruction. He is author of Standards of English in Higher Education: Issues, Challenges and Strategies (Cambridge University Press).

Gerard Sharpling is a Senior Teaching Fellow within the Centre for Applied Linguistics, University of Warwick. He has nearly 30 years of experience of working in language teaching, assessment and teacher education in the Further and Higher Education sectors. He has worked for the Open University (West Midlands and London regions) and the Faculté des Lettres et Sciences Humaines, Université de Nantes (France), as well as at the University of Birmingham. His main current research interests are in language learning, testing and assessment.

What traits do academics value in student writing? Insights from a psychometric approach

Neil Murray* and Gerard Sharpling

Centre for Applied Linguistics, University of Warwick, Coventry, UK

The number of students studying for university qualifications through the medium of English and for whom English is not their first language has increased significantly in recent years. This, along with efforts to widen access to those traditionally under-represented in higher education, has brought into focus the question of what academics see as constituting a 'good' piece of student writing. In this small-scale pilot study, Thurstone's (1959) method of paired comparisons was used to establish a scale for ranking six essays in terms of how favourably each was viewed by academic lecturers when compared with every other essay in the set. Kelly's (1955) repertory grid technique was subsequently applied to interviews conducted with the same lecturers to establish which traits they associated with the upper and lower end of the scale. Findings suggest that this methodology represents a promising approach to establishing what academic tutors regard as the key elements of good writing and thus provides an indication of what English teachers might fruitfully focus on in their teaching of the skill.

Keywords: writing assessment, rater-generated constructs, perceptions of good writing, writing for academic purposes

Introduction

English-medium universities today are increasingly characterised by a diverse student demographic as they seek to internationalise and spread their brands and influence abroad,

* Corresponding author. Email: N.L.Murray@warwick.ac.uk

and to widen participation at home. One key issue resulting from these developments is the need for students to reach a level of English language proficiency that will enable them to engage with their degree studies, reach their full academic potential and successfully graduate from their programmes (Arkoudis, Baik, and Richardson 2012; Murray 2016; Cross and O'Loughlin 2013; UK Quality Assurance Agency 2009). Writing competence is seen as particularly critical, given that it remains the predominant medium through which students are assessed during their academic careers.

Universities have traditionally relied on two (broad) mechanisms in order to ensure that students have the requisite writing skills for degree study: first the specification of language conditions of entry based on performance on internationally recognised English language gatekeeping tests such as IELTS, TOEFL and PTE and/or other pathways, such as foundation and pre-sessional programmes (Banerjee and Wall 2006; Seviour 2015); and second, the provision of in-sessional language support for those students who require it. For these mechanisms to be effective, they need to be informed by the perceptions and expectations of 'good writing' of those academics responsible for assessing student performance on course work. However, there is considerable evidence in the research literature that university lecturers are inconsistent and unreliable in their assessment of student work. While this variability has been ascribed to differing levels of professional knowledge, experience and values (see, for example, Read, Francis, and Robson 2005; Smith and Coombe 2006), empirical studies suggest that it can also be a product of assessors attaching importance to different qualities or traits in student work (O'Hagan and Wigglesworth 2014; Read, Francis, and Robson 2005; Smith and Coombe, 2006). Those qualities will often reflect implicit assessment criteria not necessarily included in formal rubrics (Baume, Yorke, and Coffey 2004; Price 2005; Read, Francis, and Robson 2005; Webster, Pepper, and Jenkins, 2000) – a fact that may go some way to explaining Bloxham, den-Outer, Hudson and Price's

observation that, taken as a whole, 'the research suggests that, even where assessors agree marks (which, in the authors' experience, lecturers often claim to), this may not necessarily be for the same reasons' (2016, 467).

While an understanding of lecturers' perceptions and expectations of good writing may not necessarily herald any imminent change to the writing components of gatekeeping tests, it can nonetheless help to guide the decisions of course designers and teachers concerning how students may be supported through writing tuition that is responsive to academics' notions of good writing and their corresponding expectations. Being aware of and able to meet those expectations is also becoming more important in light of the substantial financial outlay a university degree now represents, and the ever more competitive post-university employment environment students are required to navigate. These factors magnify the importance of good essay marks and course grades for they raise expectations on the part of students regarding their own achievement (and who increasingly see themselves as consumers or clients), and they have fuelled the need for greater explicitness in terms of how their written work will be evaluated (e.g. McSweeney 2014). Information about what academic lecturers value in written assignments thus promises to be highly beneficial to students seeking to optimise their academic performance and job opportunities.

The purpose of the study

The study we report on here was an attempt to pilot a methodology combining Thurstone's method of paired comparisons and Kelly's repertory grid technique, two methods that were used to (a) establish a scale for ranking a sample set of essays, and (b) identify which writing traits emerged as most salient for raters in relation to essays appearing at different points on the established scale. Previous studies of what constitutes a 'good' essay have been

conducted through interviews with academic tutors (e.g. Vardi 2000; Nesi 2006); marking exercises (O'Hagan and Wigglesworth 2014); combined interviews and observations (e.g. McCarthey and Mkhize 2013); and self-assessment (e.g. Orsmond et al. 1997). Kelly's repertory grid method has been used previously to analyse essay rating procedures (see Bloxham et al. 2015); however, our methodology appears to make a novel and original contribution to the body of studies in this area in its use of the combined psychological research of both Kelly (1955) and Thurstone (1959).

The aims of the study were twofold:

- to identify those writing traits that served to differentiate higher quality from lower quality assignments according lecturers' perceptions, and which might in part guide subsequent decisions regarding the focus of writing tuition; and
- to allow for the piloting of an innovative methodology to ascertain underlying critical writing constructs and the relative importance assigned to them by raters in their assessment of academic essays.

Methodology

The Method of Paired Comparisons and the Law of Comparative Judgement

This study drew, in part, on a methodology first used by Pollitt & Murray (1996) in relation to rater behaviour in the assessment of *oral* proficiency and involving the bringing together of Thurstone's method of paired comparisons (1959) and Kelly's Personal Construct Psychology (1955, 1977), both of which utilize the context of comparison. The method of paired comparisons operationalises Thurstone's Law of Comparative Judgement and, in the words of Pollitt and Murray,

> … constitutes a methodological response to the question of how to scale complex attitudes toward stimuli, and, in the words of its founder, illustrates that 'psychophysical experimentation is no longer limited to those stimuli whose physical magnitudes can be objectively measured (ibid., p. 5). Its focus is therefore *psychological* measurement; its aim, the creation of *psychological* continua upon which such stimuli may be mapped, thereby providing a means of measuring their relative values according to the degree to which subjects perceive them as encapsulating whatever attribute the researcher is concerned with (1996, 78).

What determine the scale in this technique are the comparative judgements individuals make when presented with the relevant stimuli – in the current study, a set of students' first-term postgraduate assignments. Each stimulus (or piece of writing) is compared with every other one and participants are required to make a quick and instinctive judgement as to which of the two appears greater or more dominant with respect to the attribute to be scaled – in this case, writing quality. In this methodology, no equality judgements are allowed; even if two stimuli are considered by a rater to be of similar quality (as may be the case when subject lecturers evaluate and grade student essays), participants are prompted to make a clear decision about which one is to be ranked higher. The position of each essay on the scale is 'derived from the pattern of its dominance in repeated comparisons with all of the other stimuli' (Pollitt & Murray ibid., 78); that is, those essays which are most frequently judged to be better examples of student writing appear toward the top end of the scale, and vice-versa, with the remainder located on the continuum somewhere in between.

Kelly's repertory grid technique is an application of personal construct psychology (Kelly, 1955), the basic premise of which is that as each individual experiences the world, they develop a repertoire of constructs that help them to make sense of it by bringing familiarity, order and thus a degree of predictability to it. Bannister and Fransella (1971) defined personal construct psychology as:

> … an attempt to understand the way in which each of us experiences the world, to understand our 'behaviour' in terms of what it is designed to signify and to explore how we negotiate our realities with others (27).

A construct, according to Kelly, is 'a way in which some things are construed as being alike and yet different from others' (1955, 105) – a notion that reflects his belief that we cannot know something unless we know what it contrasts with. There can be no good without an awareness of what is bad. Thus, 'Tokyo is safe whereas New York is not' indicates that safety is a construct that has significance or salience for the speaker/writer in that it helps them to organise their experience of cities, a single set of which Tokyo and New York are both members but with different characteristics that distinguish them. That is, they think about cities, in part, in terms of their degree of safety, in the same way that they may also think about them in terms of the extent to which they are cosmopolitan (another construct). In other words, our repertoire of constructs associated with cities is one that indicates the basis on which we differentiate between them.

Context and sample

The study was conducted within the researchers' university department in the UK. The department delivers a BA programme in Language, Culture and Communication, an MA in Teaching English to Speakers of Other Languages (TESOL), and an MSc in Intercultural Communication for Business and the Professions. It also offers PhD supervision and is responsible for providing English language development courses for students enrolled on award-bearing university courses, as well as external students.

Students studying on the above-mentioned MA programmes were invited to submit two essays for evaluation written during their first term of study. Both essays had been assessed in their respective programmes. University ethics guidelines were strictly adhered to in the collection and handling of the data, and informed consent was obtained from participants, who were given reassurance that having their essays evaluated for the purposes of the study would in no affect the marks that had already been awarded for the work within the context of the MA programme. Anonymity was guaranteed, with each participant given a pseudonym and each essay a code number.

A total of six students submitted pairs of essays for the purposes of the study. It was decided that, due to the length of time it was likely to take participants to reads through, comment on and evaluate all twelve pieces of writing multiple times in the interviews (see 'Procedure', below), the sample should be reduced to six essays. The essays covered a range of topics within the broad discipline of Applied Linguistics and varied in length from 1,500 words to 3,000 words (see Table 1). They were written by students for whom English was not their first language.

INSERT TABLE 1 ABOUT HERE

Although the resulting *n* size was small, it was nonetheless seen as sufficient for the purpose of trialling the methodology and obtaining an initial sense of rater behaviour in writing assessment. It thus served the purpose of what was essentially a pilot study conducted in the expectation that it would be replicated on a larger scale in the future.

Five subject lecturers participated in the study, all of whom had been involved in the delivery of content-based Applied Linguistics modules, although the essays under consideration were not necessarily related to their own particular individual specialisms. There was no minimum level of experience required for lecturers to participate in the study; however, all participants had between 1 and 28 years' experience of teaching in higher education, and between 1 and 25 years of academic lecturing experience. The difference between years of higher education experience and years of academic lecturing experience may be explained by the fact that some of the lecturers had previously occupied other teaching roles, such as language teaching positions, within Higher Education institutions. For the purposes of this study, we take 'academic lecturer' to mean a lecturer who is involved in delivering content-based teaching sessions and assessing students on their written performance on essays related to the content-based module they are delivering.

The experience of the participating lecturers is shown in Table 2.

INSERT TABLE 2 ABOUT HERE

Procedure

The five lecturers were individually asked to compare each of the six scripts with every other script in the sample, so as to cover every possible combination, and then to make an immediate holistic judgement as to which of the pair they felt was better, as per Thurstone's methodology. They were *not* asked to focus specifically on language, but neither were they discouraged from doing so. As previously stated, it was decided to limit to six the number of scripts used, due to the time-consuming nature of the task and the fact that six was deemed sufficient to provide an initial indication of the viability of the methodology and of its potential to provide insight into the qualities associated with good writing and which might, therefore, indicate a fruitful focus for teachers of academic writing. Six essays meant that each lecturer was required to read every essay five times (making a total of 30 readings) and to make 15 judgements as follows:

(a) 1 vs 2 (e) 2 vs 3 (i) 3 vs 4 (l) 4 vs 5 (n) 5 vs 6

(b) 1 vs 3 (f) 2 vs 4 (j) 3 vs 5 (m) 4 vs 6

(c) 1 vs 4 (g) 2 vs 5 (k) 3 vs 6

(d) 1 vs 5 (h) 2 vs 6

(e) 1 vs 6

The total number of comparisons made by the six raters collectively was thus 75 (15 judgements x 6 raters). This was a lengthy process, but as participants' familiarity with the six essays increased, so the comparative process became quicker. In order to ensure that their judgements were as objective as possible, participants were not made aware of the original marks awarded for the essays.

Immediately after each judgement, teachers were asked to suggest how the two scripts of each pair were qualitatively different, focussing on all aspects of the performances. Observing Pollitt and Murray's cautionary note, it was emphasised to the teachers that they should make qualitative comparisons rather evaluative ones, the idea being that those constructs integral to Kelly's personal construct psychology and underlying the quantitative Thurstonian judgements should emerge naturally.

Method of analysis

The number of times a given essay was rated more highly by raters across all the comparisons was aggregated and tabulated. This information enabled a scale to be established on which each of the six essays was positioned based on the number of times it was rated better, calculated as an overall percentage. The notes from the rater interviews were then read and re-read carefully to identify the main emerging positive and negative traits attributed by the raters. The key statements made by the raters were isolated and then categorised according to emerging, semantically-determined overarching constructs. For example, 'reading', 'referenced' and 'sources' were 'grouped' under the same construct 'Use of sources'. Instances arose, however, where the coding of the interview notes proved to be a more challenging process. For instance, 'sense of knowledgeability' could equally have been coded under both 'Use of sources' or 'Persuasive/credible', since appearing to be 'knowledgeable' can also depend on effective use of referencing. By the same token, 'sense of engagement with the topic' could have been coded under 'Persuasive/credible' or 'Identifiable voice'. Likewise, discussion took place as to whether 'use of the literature' should be coded under 'Use of sources' (as was ultimately decided) or within a new construct, 'Content use/knowledge'.

Results

Based on the interviews with raters, a table (Table 3) was created to show the comparisons made between the essays, with entries in the table referring to the number of times a given essay was cited as being the stronger of the two essays in the fifteen pairings.

INSERT TABLE 3 ABOUT HERE

Using this information gleaned from the paired comparisons, the following scale was drawn up in order to rank the six essays according to the frequency with which each essay was cited as the stronger when compared with the other five essays in the set:

INSERT FIGURE 1 ABOUT HERE

As these results indicate, the intervals between the six essays were quite distinct despite the fact of the students being in the same year and at the same stage of learning in their academic courses. From an analysis of raters' comments, it became clear that there were certain traits that most or all of the raters associated with the upper end of the writing proficiency scale and which served as key performance-differentiating factors; conversely, the essays in which these traits were less evident generally tended to be disfavoured in the performance comparisons, and were thus more associated with the lower end of the scale. These most salient positive traits are summarised in Table 4 below according to the frequency with which they were mentioned, along with sample illustrative comments that emerged during the interviews.

INSERT TABLE 4 ABOUT HERE

While, as one might expect, essays that were less favoured in the comparative judgements were often seen as lacking in the above qualities, there were a number of other discrete traits that were widely associated with the weaker essays. These included:

- poor introductions;
- a tendency to be overly formulaic;
- vagueness of language and ideas;
- an insufficient sense of audience and of what should be treated as given and new information;
- limited and imprecise vocabulary;
- lack of relevance;
- failure to juxtapose ideas;
- inappropriate register; and
- weak conclusions.

Although other traits emerged in the interviews with raters, the traits recorded here are those which were highlighted most consistently across raters in the 75 comparisons that were conducted in total.

A number of intriguing observations arose from interviews with the lecturers. It was striking, for example, how few comments explicitly focused on essay content and the demonstration of knowledge, although this trait might, arguably, be seen as encapsulated within the constructs of *persuasive/credible* and *a sense of learning on the part of the reader*. We consider this further in the discussion section.

Since raters were not being asked to mark the essays, but rather to identify the traits they noticed in each essay which differentiated them from the other five with which they

were paired, it became evident that the level of accessibility of the writing – reflected in the themes of *coherence/good flow, good argument structure, focused, and progression from theoretical to practical* – was a key factor in determining how well an essay was evaluated by the raters. It is likely that this was at least in part a product of the fact that the raters were asked to read quickly through a large volume of student writing, as is typically the case for lecturers when marking large batches of assignments, and as such accessibility is perceived as an important strength because it enables markers to complete their assessment more quickly, efficiently and with greater ease.

Discussion

It is noteworthy that the raters were not obviously preoccupied with more formal aspects of language such as accuracy and range of vocabulary, grammar and syntax, but instead were often more attentive in much of their commentary to other dimensions of the students' writing. For example, in evaluating the weakest essay (essay 2) both raters 1 and 2 indicated the main weakness as being that the writing lacked a sense of 'context' and 'explanation of purpose'. Meanwhile, these same raters praised the highest rated essay (essay 6) in terms of its ability to 'construct an argument' and 'use a theory without regurgitating it' (rater 2), and for the impression it created of being 'stylistically readable, with a thread going through it'. This suggests that academic writing programmes such as pre-sessional courses and credit- and non-credit-bearing in-sessional courses should perhaps focus less on language *per se* and more on other aspects of writing such as students' understanding of subject-specific content and their critical engagement with it. For those pre-sessional courses that run over the summer immediately preceding students' transition onto their degree programmes (rather than year-round), there is, anyway, a recognition that structural infelicities around grammar

and syntax are unlikely to be very usefully addressed in the short term and that it is more productive, therefore, to focus on other areas of writing such as accessibility (e.g. organisation), register, disciplinary genres, referencing and critical thinking. In this respect, traits identified in Table 4 that might create useful foci for writing tuition include the writer's 'voice', their ability to sustain a critical analysis and the ability to 'interact' with the literature.

In our findings, then, we observed a general tendency for subject lecturers to be less preoccupied with the individual elements of language accuracy over and above normal, summative judgements as to whether a student's language falls within acceptable parameters. In this study, one might have expected applied linguistics lecturers to take greater interest in the linguistic aspects of the written performances than academics working in other disciplines. However, our evidence seems to suggest otherwise; and other facets of the performances (e.g. remaining critical, persuasive and focused in approach) appear to take precedence and constitute more important implicit assessment criteria. At the same time, the lower level of interest in the mechanics of language use does not necessarily imply that the raters placed greater emphasis on the content of the essays; indeed, as has already been mentioned, the raters appeared to focus their immediate attention more on the macro elements of writing such as overall structure and organisation, as well as on the degree of confidence, engagement and explicitness manifested in the students' writing, and this seemed to be the case regardless of the topic under discussion.

There are two points of caution that warrant mention regarding the writing proficiency scale that emerged (Table 3). Firstly, while these traits were the most frequently cited, they do not in themselves indicate what constitutes a good essay according to any objective measure, but rather, what is salient to raters and therefore, by extension, what they perceive as important and key differentiators of quality in written performances – and thus likely part-

determiners of grades awarded, we suggest. This raises the question of the extent to which findings such as these, based as they are on academics' perceptions, align with objective statements concerning standards of good writing found in textbooks, departmental/course handbooks, test proficiency-level descriptors and documents such as the Common European Framework of Reference for Languages (CEFR). This suggests an interesting avenue for further research. Secondly, few essays manifested all those traits associated with the upper end of the scale and none of those associated with the lower end, and this raises the question of the relative weighting that different raters intuitively give to different traits. For example, would an essay which used sources effectively but lacked the author's voice be perceived more favourably than one which expressed the author's voice but used sources less effectively? These questions notwithstanding, the distance between each point in the scale remains significant, with an interval of approximately 17% between the highest and lowest rated essay.

The findings also suggest that raters tend to notice different elements within the same piece of coursework. For example, in considering the lowest rated essay (essay 2) with the highest rated essay (essay 6), the first rater stated that they preferred the higher ranking writer's 'stylistic approach', 'sense of being an expert' and being 'well-informed'. The second rater, meanwhile, praised the higher rated writer's 'sense of purpose', 'awareness of context' and 'argument'. This type of variation is certainly consistent with previous findings (e.g. O'Hagan and Wigglesworth 2014; Read, Francis, and Robson 2005; Smith and Coombe 2006; Bloxham, den-Outer, Hudson and Price 2016; and others) and may have potentially significant implications in that, in summative assessments of students' work, markers may be influenced by different traits in a given piece of work and this might affect the mark they award. Moreover, the salience of different traits for different raters suggests that attempts to achieve inter-rater reliability may be problematic (O'Hagan and Wigglesworth, ibid.). This

variability may be even more noticeable where students are writing about different topics, a fact which can serve to militate against equitable judgements in assessment (Bachman, as cited in JALT 2003).

One intriguing question concerns whether and to what extent the implicit criteria an assessor invokes, whether consciously or unconsciously, override those formal criteria frequently provided by departments to assist in the assessment of student written work and help ensure inter-rater reliability.

Limitations

Given the small-scale nature of the research and the sample size, the study we have reported on had a number of potential limitations. Confining the study to the researchers' department may have compromised the generalisability of the findings. Furthermore, the fact that the essays forming the data set were written by students in the same year and studying the same subject could also have led to only minor differences being observed between the essays. The possible effect of a fatigue factor associated with reading a large quantity of student writing in one sitting must also be acknowledged and may have had some impact on intra-rater reliability.

Nonetheless, it was felt that these potential shortcomings were outweighed by the following considerations:

- Having participants from a single department that was also the researchers' own department helped to maintain control of the project by ensuring easier access to participants. It also provided a more consistent picture of practice across one department, given that the sample was small.

- It was felt that unless the scale of the project had been *considerably* larger in terms of participant numbers and their spread across different discipline areas, extending the study beyond the department concerned would have been unlikely to increase significantly the generalisability of any findings.
- The 'intensive' design of the project in terms of the quite considerable time commitment required of participants and researchers meant that it was not practicable to increase the scale of the study, particularly in light of the fact that one of its prime objectives was to trial the effectiveness, in a new skill area (writing), of the two-pronged methodology only previously employed to explore rater behaviour in the assessment of oral proficiency.
- It was felt that far from compromising intra-rater reliability, involving the raters in an intensive assessment of six essays in one sitting might actually result in greater intra-rater reliability, since (a) there was less opportunity to 'forget' key issues between separate evaluation periods, and (b) reading efficiency developed quickly once the raters had got their 'eye in' and increased their familiarity with the scripts.

Conclusion

Overall, this study has sought to examine raters' views of the academic writing of a small sample of students. Given that the study was small-scale, involving participants from only one department of the University, the generalisability of its findings is necessarily limited. The traits identified in Table 3 are relatively broad, and lack the more subtle, fine-grained nature of the qualitative written (and spoken) feedback that is generally given to students. The researchers were also conscious of the fact that, with participants being applied linguists, there may have been an unrepresentative tendency for them to focus disproportionately more

on language quality (e.g. accuracy and range of vocabulary and structures) than on other aspects of students' written performances, despite instructions to raters at the outset to consider all aspects. This assumption was disproved to some extent by our findings, which suggest that other traits were regarded as more important; nevertheless, we have to compare our findings from this study with traits identified by lecturers in other subject areas.

This study appears to show a considerable degree of consistency of judgements across raters, as evidenced by a commonality of salient constructs that emerged and perceptions of their relative importance, and by the scale that emerged during the analysis of the data. These results highlight the potential contribution that can be made by using Thurstone's method of paired comparisons and Kelly's repertory grid technique in combination, and as originally applied to the assessment of oral performance by Pollitt and Murray (op. cit.) to evaluate students' oral performances. Despite the modest scale of the study, the findings suggest that the application of these methods can foreground revealing issues concerning academics' perceptions and expectations of student writing, and this information might be of value to students seeking to gain higher marks in their assignments with a view to acquiring strong degrees that will position them advantageously in a global world context where competition for jobs and research awards is increasingly fierce.

Further analysis is needed to ascertain what traits are regarded as important or most valued in particular kinds of writing in a range disciplines, in order to establish whether broad cross-disciplinary traits are identifiable and to better ensure that greater standardisation in marking fosters more secure statements about where and how students improve in their writing skills (Sharpling 2002). Our findings could, for example, be cross-referenced with descriptions found in the literature of what is acceptable writing practice in specific disciplines (e.g. Nesi and Gardner 2012). It would also be beneficial to undertake further research to ascertain the views of students regarding the same essays evaluated by the

lecturers, and to elicit evaluations from language tutors who are involved in preparing students for academic study. The methodology adopted in this study suggests a positive way forward in engaging in dynamic qualitative research on academic writing within and across disciplines.

References

Arkoudis, S., C. Baik, and S. Richardson. 2012. *English Language Standards in Higher Education*. Camberwell, Victoria: ACER (Australian Council for Educational Research) Press.

Banerjee, J., and D. Wall. 2006. "Assessing and Reporting Performances on Pre-sessional EAP Courses: Developing a Final Assessment Checklist and Investigating its Validity." *Journal of English for Academic Purposes* 5 (1): 50–69.

Bannister, D., and F. Fransella. 1971. *Inquiring Man: The Psychology of Personal Constructs*. London: Croom Helm Ltd.

Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016. "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria." *Assessment & Evaluation in Higher Education* 41 (3): 466–481.

Cross, R., and K. O'Loughlin. 2013. "Continuous Assessment Frameworks within University English Pathway Programs: Realizing Formative Assessment within High-Stakes Contexts." *Studies in Higher Education* 38 (4), 584–594. doi:10.1080/03075079.2011.588694.

JALT (Japan Associated of Language Teachers). 2003. Voices in the Field: An Interview with Lyle Bachman. *JALT Testing & Evaluation SIG Newsletter*, 7 (2): 12–15). Accessed April 9 2018. http://hosted.jalt.org/test/bac_hub.htm

Kelly, G. A. 1977. “The Psychology of the Unknown.” In *New Perspectives in Personal Construct Theory*, edited by D. Bannister, 1–19. London: Academic Press.

Kelly, G. A. 1955. *The Psychology of Personal Constructs, Volumes I and II*. New York: Norton.

McCarthey, S. J., and D. Mkhize. 2013. “Teachers' Orientations Towards Writing.” *Journal of Writing Research* 5 (1): 2–33.

McSweeney, F. 2014. “Students' views on assessment.” *Other resources 12* [Online]. Accessed March 28 2018. https://arrow.dit.ie/aaschssloth/12

Murray, N. 2016. *Standards of English in Higher Education: Issues, Strategies and Challenges*. Cambridge: Cambridge University Press.

Nesi, H., and S. Gardner. 2006. “Variation in Disciplinary Culture: University Tutors' Views on Assessed Writing Tasks.” *British Studies in Applied Linguistics*, no. 21: 99–117.

Nesi, H., and S. Gardner. 2012. *Genre Across the Disciplines: Students’ Writing in Higher Education*. Cambridge: Cambridge University Press.

O'Hagan, S., and G. Wigglesworth. 2014. “Who's Marking my Essay? The Assessment of Non-Native-Speaker and Native-Speaker Undergraduate Essays in an Australian Higher Education Context.” *Studies in Higher Education* 40 (9): 1729–1747.

O’Loughlin, K. 2009. “‘Does it Measure Up?’ Benchmarking the Written Examination of a University English Pathway Program.” *Melbourne Papers in Language Testing* 14 (1): 32–54.

Orsmond, P., S. Merry, and K. Reiling. 1997. “A Study in Self-Assessment: Tutor and Students’ Perceptions of Performance Criteria.” *Assessment & Evaluation in Higher Education* 22 (4): 357–368.

Pollitt, A., and N. Murray. 1996. “What Raters Really Pay Attention to.” In *Performance Testing, Cognition and Assessment: Selected Papers from the 15th. Language Testing*

Research Colloquium (LTRC) — Cambridge and Arnhem, 1993, edited by M. Milanovic and N. Saville, 74–91. Cambridge: Cambridge University Press.

Quality Assurance Agency for Higher Education (QAA). 2009. *Thematic Enquiries into Concerns about Academic Quality and Standards in Higher Education in England*. Gloucester: Quality Assurance Agency for Higher Education. Accessed February 9 2018. http://www.qaa.ac.uk/Publications/.../Documents/FinalReportApril09.pdf

Read, B., B. Francis, and J. Robson. 2005. "Gender, Bias, Assessment and Feedback: Analyzing the Written Assessment of Undergraduate History Essays." *Assessment & Evaluation in Higher Education* 30 (3): 241–260. doi:10.1080/02602930500063827.

Seviour, M. 2015. "Assessing Academic Writing on a Pre-sessional EAP Course: Designing Assessment which Supports Learning." *Journal of English for Academic Purposes*, no. 18: 84–89. doi:10.1016/j.jeap.2015.03.007.

Sharpling, G. 2002. "*Learning* to Teach English for Academic Purposes: Some Current Training and Development Issues." *English Language Teacher Education and Development*, no. 6: 82–94.

Smith, E., and K. Coombe. 2006. "Quality and Qualms in the Marking of University Assignments by Sessional Staff: An Exploratory Study." *Higher Education* 51 (1): 45–69. doi:10.1007/s10734-004-6376-7.

Thurstone, L.L. 1959. *The Measurement of Values*. Chicago: University of Chicago Press.

Vardi, I. 2000. "What Lecturers Want: An Investigation of Lecturers' Expectations in First Year Essay Writing Tasks." In Ponencia presentada en la Forth Pacific Rim, First Year in Higher Education Conference. [Online]. Accessed March 24 2018. http://fyhe.com.au/past_papers/papers/VardiPaper.doc